Fraud Detection with Machine Learning in Property Insurance Policy Requests

Sergül Ürgenç[1*], Habip Kaplan[2], Ayça Çakmak Pehlivanlı[3]

[1]Jforce Information Technologies, Istanbul, Turkiye, sergul.urgenc@jforce.com, ORCID: 0000-0003-1965-4488*

[2]Jforce Information Technologies, Istanbul, Turkiye, habip.kaplan@jforce.com, ORCID: 0000-0003-3512-8955

[3]Mimar Sinan Fine Arts University, Istanbul, Turkiye, ayca.pehlivanli@msgsu.edu.tr, ORCID: 0000-0001-9884-6538

The purpose of this study is to predict in advance, whether the policy claims are made for abuse in order to reduce claim payments in the property branch of the insurance industry and to prevent any financial losses. In order to detect abnormal situations, Label Spreading and Self Training semi-supervised machine learning approaches were used due to the insufficient label ratio of the anomaly class in the data set. After the data labeling process, fraud prediction study was conducted with supervised machine learning models and it was discussed which semi-supervised learning approach worked with higher performance. The datasets include the policy demands in 2017 and 2018 and the weather information of that location. Accuracy, precision, recall, specificity, and F1 score were evaluated as model success measures, and according to these results, it was seen that the Self Training approach could label data with higher performance than the Label Spreading approach.

Keywords: Fraud Prediction, Machine Learning, Property Insurance, Semi Supervised Learning.

1. Introduction

Although “insurance fraud” occurs in the insurance branch, it is mostly experienced in automobile, fire and health insurances. The significant increase in financial losses compared to the increase in premiums over the years is also noteworthy in the fire branch. Conducting fraud detection with only expert knowledge without analytical studies is insufficient with the increasing data volume. It is desired to prevent a policy from being requested when there is a possibility of damage or actual damage occurs without having a policy.

Due to the large data volume, detecting and labeling fraud data manually is a difficult process. The tagged data rate may be very low or nonexistent. Data that is not labeled as fraudulent and considered to be normal may contain policy requests that are actually made for fraudulent purposes. Ensuring the accuracy of the data is a difficult step and a time-consuming process, as it directly affects the success of the prediction models and due to the large volumes of data. This may cause the model to learn incorrectly and decrease its success rate.

Although there have been extensive studies on fraud detection in insurance for years, studies on increasing the rate of labeled data with semi-supervised learning have started to be made in the last few years. In a study based on semi-supervised learning methods, insurance fraud data from 2015-2016 were used. A new clustering score criterion is introduced to address the problem of datasets with skewed data and large numbers of unlabeled data. The main goal is to increase the number of correctly detected fraudulent claims and reduce the proportion of genuinely non-fraudulent claims [1]. Three feature selection algorithms were applied to the large dataset containing automobile insurance damages; tree-based, L1-based and univariate.

Random Forest, Naive Bayes, K Nearest Neighbor and Decision Tree were preferred as classification algorithms in the study. When all the models were compared, it was concluded that the Random Forest model was the best model in terms of precision and sensitivity [2]. The results showed that Deep Neural Network and ensemble-based approaches such as Random Forest and Gradient Boosting gave the best success and outperformed algorithms such as Logistic Regression which are widely used in the field. In addition, the profile of the approved fraudsters in the dataset was analyzed and the effect of each feature on the overall classification performance was examined, and explanatory artificial intelligence methods were used for False Positive and False Negative measures [10]. Three different machine learning models have been developed and tested to classify Medicare provider claims. A fraud detection study was conducted using 2013 and 2016 Medicare durable medical equipment damage data to provide meaningful comparative analyzes between model evaluation methods. The models trained in this study were Gradient Boosted Trees, Logistic Regression and Random Forest. As a result of the comparison of the differences in the average AUC scores, it was seen that the training and testing method was 0.05347 points higher than the cross-validation method. ANOVA and Tukey's HSD tests were used to understand whether there was a significant difference between these two results. It was concluded that there was no significant difference between the two evaluation methods in big data [5]. A fraud detection study was carried out in the motor insurance branch. K Nearest Neighbor, Decision Tree, Logistic Regression and Neural Network algorithms have been tested in model creation. According to the results obtained in the research, Logistic Regression, Decision Tree and Neural Network models trained with 15 independent variable data sets gave more accurate results. With the dataset including 20 independent variables, better results were obtained only in the Neural Network model [11]. Two-class machine learning methods were used to automatically detect and prevent card transaction fraud. Principal Component Analysis was used to reduce the data size. In the study, Logistic Regression, Gradient Boosted, CS Logit and CS Boost models were compared. It has been seen that the newly proposed CS Logit and CS Boost models are models that minimize financial losses due to abuse [8]. A new deep learning methodology has been proposed in order to detect fraudulent behavior of the insured. Autoencoder and Variational Autoencoder deep learning methods were applied and their underlying dynamics were analyzed. A methodology is proposed that will enable these methods to be used in variable importance analysis for supervised and unsupervised learning. As a result, it was seen that the proposed unsupervised deep learning variable importance methodology provides outstanding performance in the absence of labeled data according to supervised variable importance [9]. In order to detect fraudulent behavior in purchasing transactions, 48 different types of fraud are modeled by means of statistical indicators. Using the outputs, a concept was designed to give points to the purchasing supervisor for each transaction. Outliers were detected with unsupervised learning. In the study, the PACE was evaluated incrementally according to the 0.5, 1.0, and 2.0 assessments. The outputs are combined with the rule engine, focusing on an inclusive approach [6]. There has been a study about identifying a type of abuse in which small telephone operators inflate the number of calls coming into their networks by taking advantage of a higher access fee than the network operator associated with the origin of the call. Due to the lack of labeled data, a decision support system has been proposed for fraud detection with clustering and Decision Tree methods. After data collection and feature engineering, DBSCAN and OPTICS unsupervised learning models and potential fraud cases are grouped into clusters. Then, the cluster memberships were examined and the data were labeled. A fraud prediction model was developed with the Decision Tree algorithm using the labeled data set [12]. Different data engineering techniques have been proposed to increase the performance of the models developed in order to understand the reasons behind the marking of a case as suspicious in banking transactions. Logistic Regression, Decision Tree, Gradient Boosted models were developed with over-sampling methods SMOTE, ADASYN, MWMOTE and ROSE and their results were compared. While Logistic Regression and Decision Tree can benefit from over-sampling methods, the overall performance of Gradient Boosted Trees has been observed to decrease. [4]. A new approach has been applied to optimize the performance of the models developed for credit card fraud detection. Due to the difficulties brought by the unbalanced data set, a new composition-based algorithm is proposed in this study, which is an approach that combines over-sampling and feature selection methods to find the best combination of several supervised classification algorithms. Ababoost, Random Forest, Multilayer Perceptron, K Nearest Neighbor, Decision Tree, Gaussian Naive Bayes models are trained and compared. A hybrid algorithm succeeded in finding the best combination of algorithms, both triplets (Borderline, Fix-RFE, Radom Forest) and (Borderline, OLS, Random Forest) algorithms has been found to have higher success than other rendering algorithms. [3]. Primitive Sub-Peer Group Analysis (PSPGA) was conducted to detect suspicious behavior in healthcare insurance. Unsupervised learning methods SmartSifter, Homogenous Subgrouping, Drift Learning, Genetic Support Vector Machines were compared with traditional supervised learning algorithms. The PSPGA's approach to detecting healthcare fraud has been shown to outperform other methods [7].

In this study, it is aimed to prevent abuses in property branch policy requests by detecting them beforehand with machine learning. The work consists of two stages. First of all, it was tried to identify and label anomalous cases in policies that are assumed to be non-abuse by using information such as coverage, channels and tariffs of the policies, with semi-supervised learning. For this purpose, Label Spreading and Self Training methods were used and the results were compared. Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB) algorithms have been developed for the Self Training method. In the second stage, fraud prediction was made with the abnormal policies labeled in the first stage. Supervised learning models were developed using weather data and they were compared according to the fraud prediction results to determine which semi-supervised learning method had better labeling. RF and C5.0 Decision Tree models have been developed for Fraud prediction. The developed supervised models were used via ensembling. Accuracy, precision, recall, specificity and F1 score metrics were used as success criteria of the models.

2. Dataset Generation

It has been studied on the masked data set belonging to the property branch from an insurance company. The datasets include the policy demands in 2017 and 2018 and the weather information of that location. In the study, a single city was selected as a sample and 194.787 policies were studied. The attributes in the datasets and their definitions are given in Table 1.

Table 1. *Attributes in the dataset.*

Dataset	**Attribute Name**	**Definition**
Policy	**Anomaly_Policy**	indicates whether policy is fraudulent (flag)
Policy	**New_Policy**	indicates whether the policy is new or renewal (flag)
Policy	**Tariff_Code**	tariff of the policy (categorical)
Policy	**Channel_Code**	information on which channel the sale was made through (categorical)
Policy	**Coverage_Code**	information on the guarantees in the policy (categorical)
Policy	**Addendum_Reason**	Includes addendum reasons (categorical)
Weather	**Num_Rainy**	number of rainy days (continuous)
Weather	**Max_Temp**	maximum temperature (continuous)
Weather	**Min_Temp**	minimum temperature (continuous)
Weather	**Mean_Temp**	mean temperature (continuous)
Weather	**Max_Temp_Avg**	maximum temperature average (continuous)
Weather	**Min_Temp_Avg**	minimum temperature average (continuous)
Weather	**Total_ Precipitation**	total precipitation (mm) (continuous)
Weather	**Num_Hail**	number of days with hail (continuous)
Weather	**Max_Wind_Speed**	maximum wind speed (continuous)
Weather	**Max_ Precipitation**	maximum precipitation (mm) (continuous)

3. Methods and Application

A. Data Preparation

Categorical variables were prepared as dummy variables to be used in semi-supervised and supervised machine learning studies. Feature selection was applied with filtering method using Pearson Correlation statistics. As a result of Feature Selection, it was deemed appropriate to use 18 independent variables for the first stage and 5 independent variables for the second stage (Table 2). The dataset was randomly divided into 80% train and 20% test, and the prediction successes of the model were compared separately.

Table 2. *Independent variables selected with Feature Selection.*

Independent variables for semi supervised learning	**Independent variables for fraud prediction**
New_Policy	New_Policy
Tariff_Code_Y58	Max_Temp
Tariff_Code_Y75	Total_ Precipitation
Channel_Code_5	Num_Hail
Coverage_Code_BN2	Max_Wind_Speed
Coverage_Code_EK3	
Coverage_Code_TP1	
Coverage_Code_SEL	
Coverage_Code_IAY	
Coverage_Code_ES1	
Coverage_Code_SL1	
Coverage_Code_EK2	
Coverage_Code_SL2	
Coverage_Code_KKB	
Coverage_Code_GRV	
Coverage_Code_BN1	
Coverage_Code_ODE	

B. Undersampling

It is the reduction of the number of dominant classes in imbalanced data to a level equal to or close to the number of lower classes. There is data loss, but it prevents synthetic data generation [4].
The data set was balanced with the undersampling method. In the data balancing study, the records belonging to the class 0 that could not be included in the sampling were considered unlabeled (Table 3).

Table 3. *Class distributions before and after balancing.*

Anomaly Policy Class	**Imbalanced Class Distribution**	**Class Distribution After Undersampling**	**Class Distribution for Label Spreading**
0	172.383	22.230	17.947
1	22.404	22.404	17.899
-1	-	-	149.979

C. Label Spreading

It is a Graph based semi supervised machine learning algorithm. It multiplies the label ratio in the data set by assigning labels to previously unlabeled data. It makes two assumptions: nearby points are likely to have the same label, and points of the same structure are likely to have the same label [13].

The hyperparameters used in the model are shown in Table 4.

Table 4. *Hyperparameter values of Label Spreading.*

Hyperparameter	Values
kernel	knn
gamma	20
n_neigbors	7
alpha	0.2
max_iter	30
tol	0.001

D. Self-Training

In the Self Training method, the model is trained using supervised machine learning algorithms with labeled data. Then this model is used to predict the unlabeled data class. In the Self Training method, the model is trained using supervised machine learning algorithms with labeled data. This model is then used to predict the unlabeled data class. Observations with a high probability of prediction are labeled and the process is iterated if necessary [16].

SVM, RF and XGB models have been developed for Self Training, hyperparameters are given in Table 5, Table 6 and Table 7, respectively.

Random Forest (RF): It is an estimator that uses a set of decision trees combined with sub-samples of the dataset using the bagging method. Each decision tree votes for the predicted class or averages the incoming predictions. The main idea of RF is to perform bagging by including only a subset of variables during tree development [15].

Table 5. *Hyperparameter values of RF.*

Hyperparameter	Values
tree method	auto
max depth	6
min child weight	1
max delta step	0.0
boost round	10
sub sample	1.0
eta	0.3
gamma	0.0
colsample by tree	1.0
colsample by level	1.0
lambda	1.0
alpha	0.0

Support Vector Machine (SVM): Generally, it is one of the supervised learning methods used in classification problems. It is defined as a vector space-based machine learning method that finds a decision boundary between the two classes that are farthest from any point in the training data. It is mostly used to separate data consisting of two classes [14].

Table 6. *Hyperparameter values of SVM.*

Hyperparameter	Values
stopping criteria	0.001
kernel	rbf
regularization	10
epsilon	0.1
rbf gamma	0.1
gamma	1.0
bias	0.0
degree	3

XGBoost Tree (XGB): XGB is a high-performance version of the Gradient Boosting algorithm optimized with various modifications. The most important features of the algorithm are its ability to obtain high predictive power, to prevent over-learning, to manage empty data and to do them quickly [13].

Table 7. *Hyperparameter values of XGB.*

Hyperparameter	Values
number of trees	10
min leaf node size	20
max depth	10
splitting method	auto
target	0.01
max iteration	50
max evaluation	100

Ensembled Model: RF, XGB, SVM are used by ensembled modeling with confidence weighted voting method. Observations that the model predicted as anomaly and false positives in the train dataset were labeled as class 1.

E. Supervised Learning for Fraud Prediction

In order to see which of the Self Training and Label Spreading approaches are more successful, fraud prediction models have been developed by using the labeled dataset and supervised machine learning algorithms. C5.0 and RF models were developed separately for both approaches, with the same hyperparameters. Hypermarameters are given in Table 8 and Table 9, respectively. Afterwards, these two models were ensembled with the confidence weighted voting method.

C5.0 Decision Tree: Tree models operate according to a set of statistical division rules aimed at maximizing the homogeneity of the dependent variable in each of the resulting groups. Each decision tree approach uses different statistical methods and a score is calculated for each group. C5.0 model works by splitting the sample based on the feature, which provides the maximum information gain [14].

Table 8. *Hyperparameter values of C5.0 for fraud prediction.*

Hyperparameter	Values
max tree depth	25
predictor impotance	true
mode	simple

Table 9. *Hyperparameter values of RF for fraud prediction.*

Hyperparameter	Values
number of models build	100
max nodes	10000
max depth	10
min child node size	5

4. Results and Discussion

The results of Label Spreading and Self Training approaches are given in Table 10. The success rate of the RF, SVM and XGB models developed for Self Training are almost the same. For this reason, the models were ensembled and used. Since the train and test success rates of the models are almost identical, only the test results are presented.

Table 10. *Test results of the semi supervised learning.*

Model	Accuracy	Precision	Recall	Specifity	F1 Score
Label Spreading	0.50	0.16	0.83	0.46	0.27
Self Training (RF)	0.70	0.67	0.77	0.62	0.72
Self Training (SVM)	0.70	0.67	0.78	0.62	0.72
Self Training (XGB)	0.70	0.67	0.78	0.62	0.72
Self Traning (Ensembled Model)	0.70	0.67	0.78	0.62	0.72

When the recall value is examined in the Label Spreading approach, the test success of the class 1 is 83%. The success of the class 0 is not enough. However, since our aim in the study was to increase the label rate of the class 1, the success of the class 0 was ignored. When we look at other success measures, it is seen that the model does not learn well enough. Policies estimated as class 1 are labeled with the Label Spreading model.

As in the Label Spreading approach, when the recall and specifity values are examined in the Self Training model, it is seen that the success rate of class 1 is higher than the success rate of class 0. However, it is seen that the Self Training model learns more in a balanced way compared to the Label Spreading approach.

Data are labeled according to Label Spreading and Self Training models. It is seen that the rate of data labeled with Self Training is more realistic. Class distributions after labeling are given in Table 11.

Table 11. *Class distributions after data labeling.*

Anomaly Class	Count (Label Spreading)	Count (Self Training)
0	82.600	114.440
1	112.187	80.347

The results of the fraud prediction test to measure which approach is more successful in the data labeling phase are given in Table 12. C5.0 and RF models are ensembled.

Table 12. *Test results of the supervised learning models for fraud prediction.*

Data Labeling Approach	Accuracy	Precision	Recall	Specifity	F1 Score
C5.0 (Label Spreading)	0.69	0.65	0.80	0.57	0.72
RF (Label Spreading)	0.67	0.65	0.72	0.63	0.68
Ensembled (Label Spreading)	0.68	0.66	0.73	0.64	0.69
C5.0 (Self Training)	0.89	0.84	0.95	0.83	0.89
RF (Self Training)	0.87	0.84	0.91	0.83	0.87
Ensembled (Self Training)	0.88	0.84	0.92	0.83	0.88

When the recall and specifity values of the ensembled models are examined, it is seen that the supervised model developed for fraud prediction achieved higher success in the data set labeled with the Self Traning method than the Label Spreading approach in classes 0 and 1. When the overall accuracy values of the two approaches were compared, the Self Training approach showed high performance with 88% accuracy. In addition, when the recall value is examined, it is seen that the anomaly class was predicted with 92% success. The difference between accuracy, precision, recall, specificity and F1 score results of both approaches were measured as 20%, 18%, 24%, 19% and 19%, respectively.

5. Conclusion

In the study, the ratio of the anomaly class, which constitutes a small part of the dataset, was replicated using semi-supervised machine learning models. While developing semi-supervised models, policy information such as coverage, tariff, and channel were used.

Policy claims for abuse were estimated by supervised machine learning models, using datasets whose label ratio was replicated separately according to both Label Spreading and Self Training model results. In the developed predictive models, weather data and whether the policy is issued for the first time are seen as important factors.

As a result of the study, it was observed that in fraud prediction the dataset labeled with the Self Training approach outperformed the Label Spreading approach, especially in accuracy, recall and specificity success metrics.

REFERENCES

[1] S. M. Palacio, “Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning,” *Data Science Journal*, vol. 18, no. 1, Art. no. 1, , Jul. 2019.

[2] S. Panigrahi and B. Palkar, “Comparative Analysis on Classification Algorithms of Auto-Insurance Fraud Detection based on Feature Selection Algorithms,” *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 72–77, Sep. 2018.

[3] I. Bouzgarne, Y. Mohamed, O. Bouattane, and Q. Mohamed, “Composition of Feature Selection Methods and Oversampling Techniques for Banking Fraud Detection with Artificial Intelligence,” *International Journal of Engineering Trends and Technology*, vol. 69, pp. 216–226, Nov. 2021.

[4] B. Baesens, S. Höppner, and T. Verdonck, “Data engineering for fraud detection,” *Decision Support Systems*, vol. 150, p. 113492, Nov. 2021.

[5] R. A. Bauder, M. Herland, and T. M. Khoshgoftaar, “Evaluating Model Predictive Performance: A Medicare Fraud Detection Case Study,” in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Los Angeles, CA, USA, Jul. 2019, pp. 9–14.

[6] A. Westerski, R. Kanagasabai, E. Shaham, A. Narayanan, J. Wong, and M. Singh, “Explainable anomaly detection for procurement fraud identification—lessons from practical deployments,” *International Transactions in Operational Research*, vol. 28, no. 6, pp. 3276–3302, 2021.

[7] L. Settipalli and G. R. Gangadharan, “Healthcare fraud detection using primitive sub peer group analysis,” *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, p. e6275, 2021.

[8] S. Höppner, B. Baesens, W. Verbeke, and T. Verdonck, “Instance-dependent cost-sensitive learning for detecting transfer

fraud," *European Journal of Operational Research*, vol. 297, no. 1, pp. 291–300, Feb. 2022.

[9] C. Gomes, Z. Jin, and H. Yang, "Insurance fraud detection with unsupervised deep learning," *Journal of Risk and Insurance*, vol. 88, no. 3, pp. 591–624, 2021.

[10] M. K. Severino and Y. Peng, "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata," *Machine Learning with Applications*, vol. 5, p. 100074, Sep. 2021.

[11] Ö. Şahin, S. Ayvaz, and E. ÇALIMFİDAN, "Sigorta Sektöründe Sahte Hasarların Tahmini İçin Geliştirilen Makine Öğrenmesi Modellerinin Kıyaslanması," *Bilişim Teknolojileri Dergisi*, vol. 13, pp. 479–489, Oct. 2020.

[12] M. E. Irarrázaval, S. Maldonado, J. Pérez, and C. Vairetti, "Telecom traffic pumping analytics via explainable data science," *Decision Support Systems*, vol. 150, p. 113559, Nov. 2021.

[13] Y. Kang, N. Jia, R. Cui, and J. Deng, "A graph-based semi-supervised reject inference framework considering imbalanced data distribution for consumer credit scoring," *Applied Soft Computing*, vol. 105, p. 107259, Jul. 2021.

[14] K. K. Tsiptsis and A. Chorianopoulos, *Data Mining Techniques in CRM: Inside Customer Segmentation*, 1st edition. Wiley, 2009.

[15] J. Wang, *Encyclopedia of Data Warehousing and Mining, Second Edition*, 2nd edition. Hershey: Information Science Reference, 2008.

[16] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, Jan. 2009.